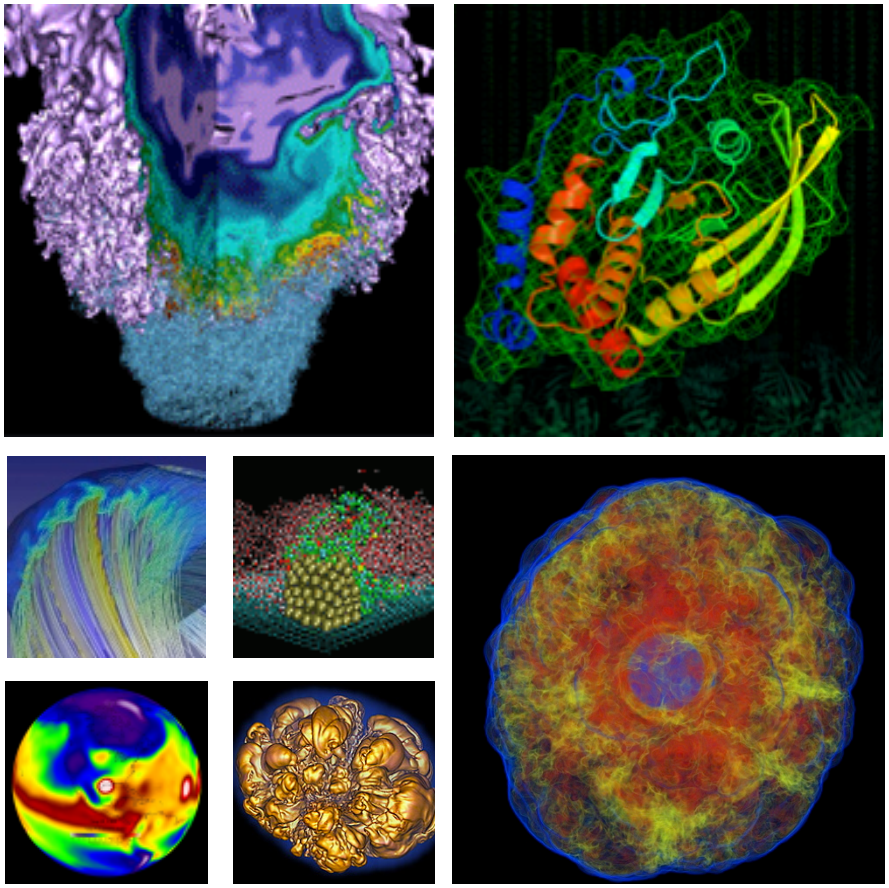
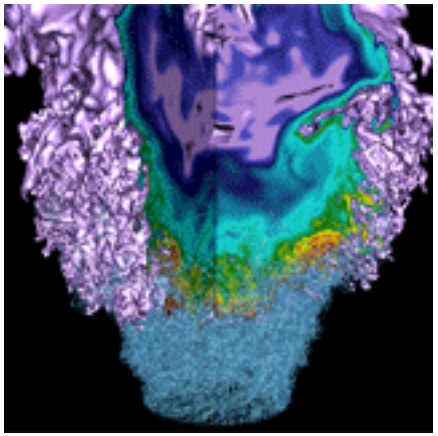
NERSC's 10 year plan

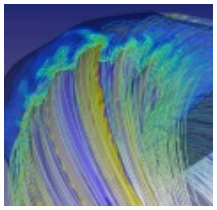


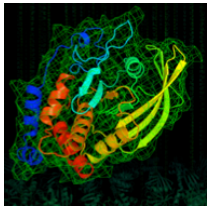
Sudip Dosanjh
Director

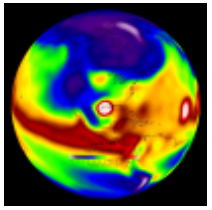
March 19, 2013

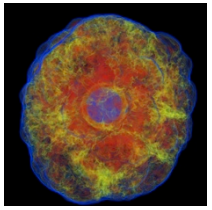
NERSC Overview

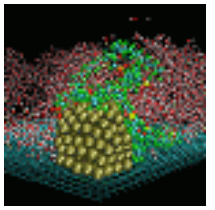




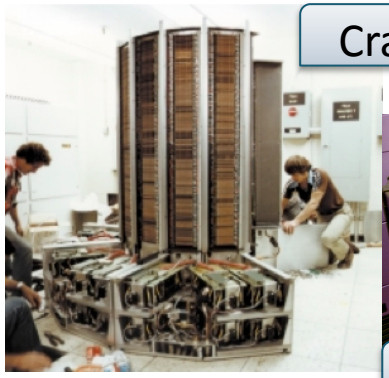




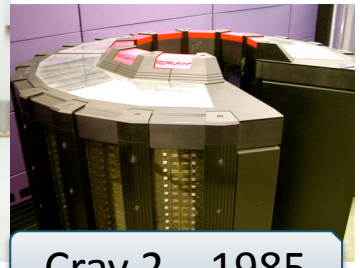




NERSC History



Cray 1 - 1978



Cray 2 – 1985

Cray T3E Mcurie - 1996

IBM Power3 Seaborg - 2001

1974	Founded at Livermore to support fusion research with a CDC system
1978	Cray 1 installed
1983	Expanded to support today's DOE Office of Science
1986	ESnet established at NERSC
1994	Cray T3D MPP testbed
1994 - 2000	Transitioned users from vector processing to MPP
1996	Moved to Berkeley Lab
1996	PDSF data intensive computing system for nuclear and high energy physics
1999	HPSS becomes mass storage platform
2006	Facility wide filesystem
2010	Collaboration with JGI

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

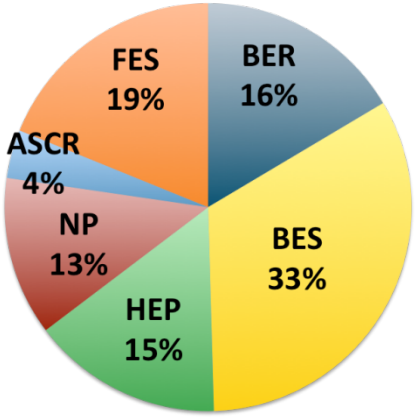
NERSC collaborates with computer companies to deploy advanced HPC and data resources

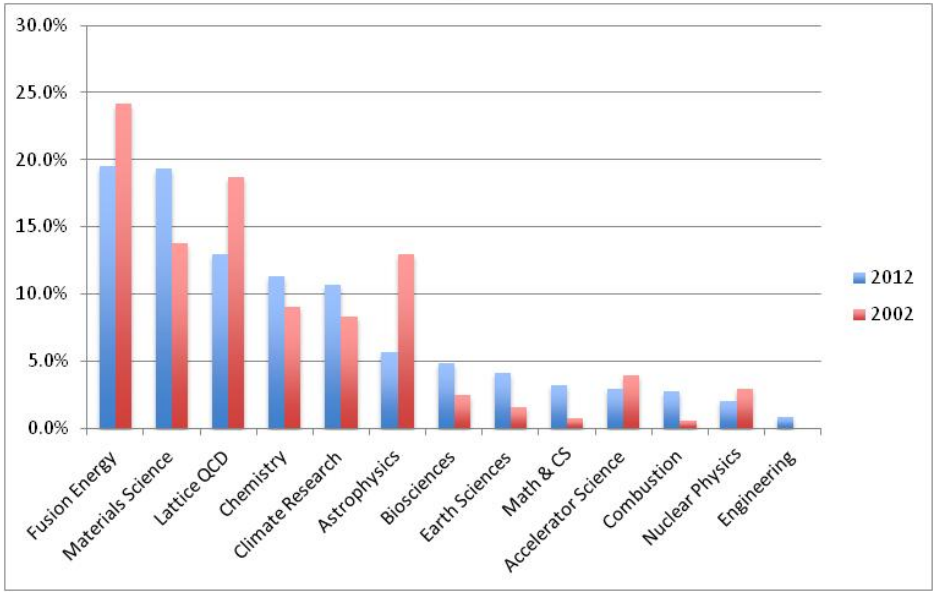
- Hopper (N6) and Cielo (ACES) were the first Cray petascale systems with a Gemini interconnect
- Edison (N7) will be the first Cray petascale system with Intel processors, Aries interconnect and Dragonfly topology (serial #1)
- N8 and Trinity (ACES) are being jointly designed as on-ramps to exascale
- Architected and deployed data platforms including the largest DOE system focused on genomics
- One of the first facility-wide filesystems

We employ experts in high performance computing, computer systems engineering, data, storage and networking

We directly support DOE's science mission

- **We are the primary computing facility for DOE Office of Science**
- **DOE SC allocates the vast majority of the computing and storage resources at NERSC**
  - Six program offices allocate their base allocations and they submit proposals for overtargets
  - Deputy Director of Science prioritizes overtarget requests
- **Usage shifts as DOE priorities change**



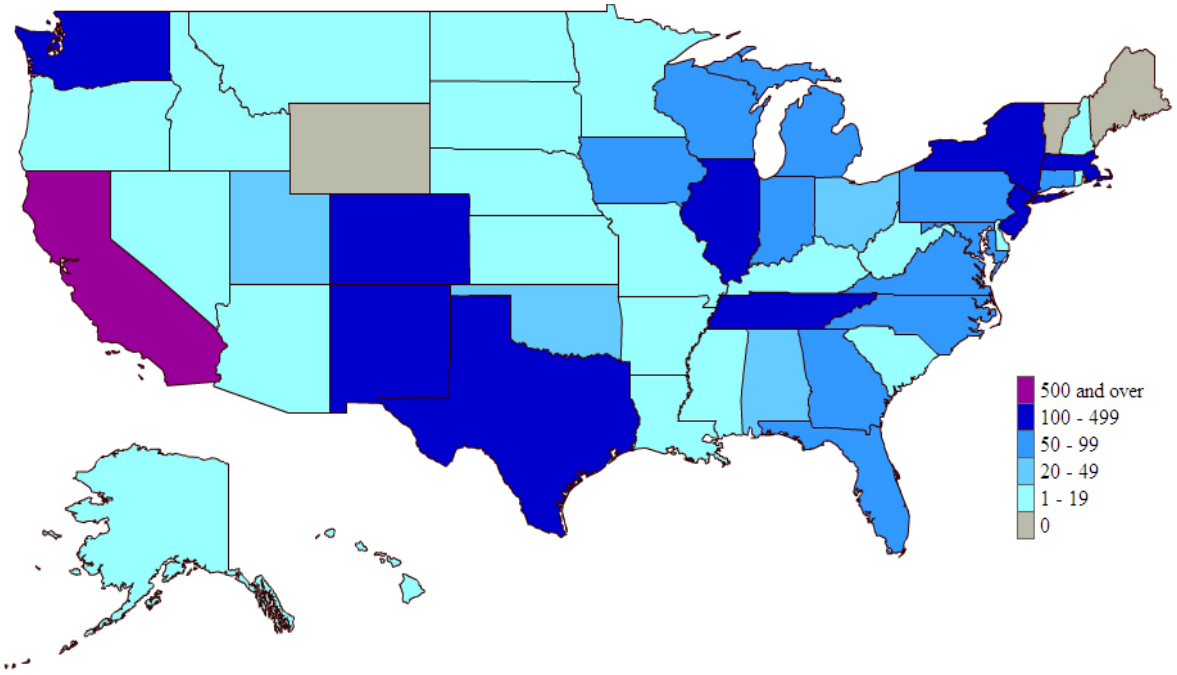


We focus on the scientific impact of our users

- 1500 journal publications per year
- 10 journal cover stories per year on average
- Simulations at NERSC were key to two Nobel Prizes (2007 and 2011)
- Supernova 2011fe was caught within hours of its explosion in 2011, and telescopes from around the world were redirected to it the same night
- Data resources and services at NERSC played important roles in two of Science Magazine's Top Ten Breakthroughs of 2012 — the discovery of the Higgs boson and the measurement of the Θ_{13} neutrino weak mixing angle
- MIT researchers developed a new approach for desalinating sea water using sheets of graphene, a one-atom-thick form of the element carbon. Smithsonian Magazine's fifth "Surprising Scientific Milestone of 2012."
- Four of Science Magazine's insights of the last decade (three in genomics, one related to cosmic microwave background)

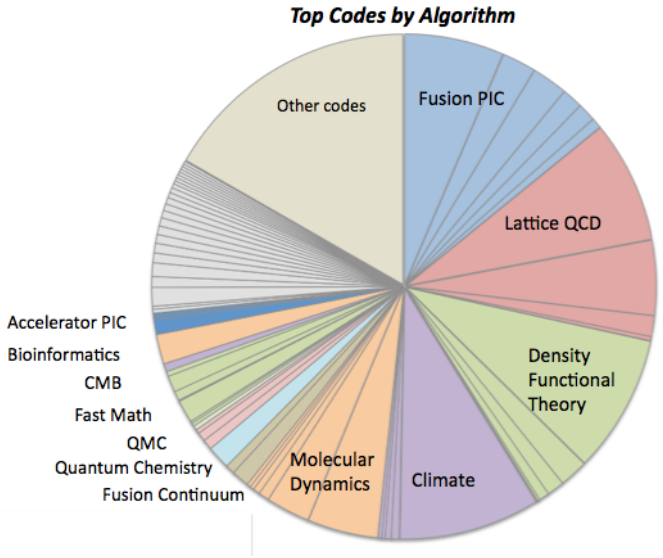
We support a broad user base

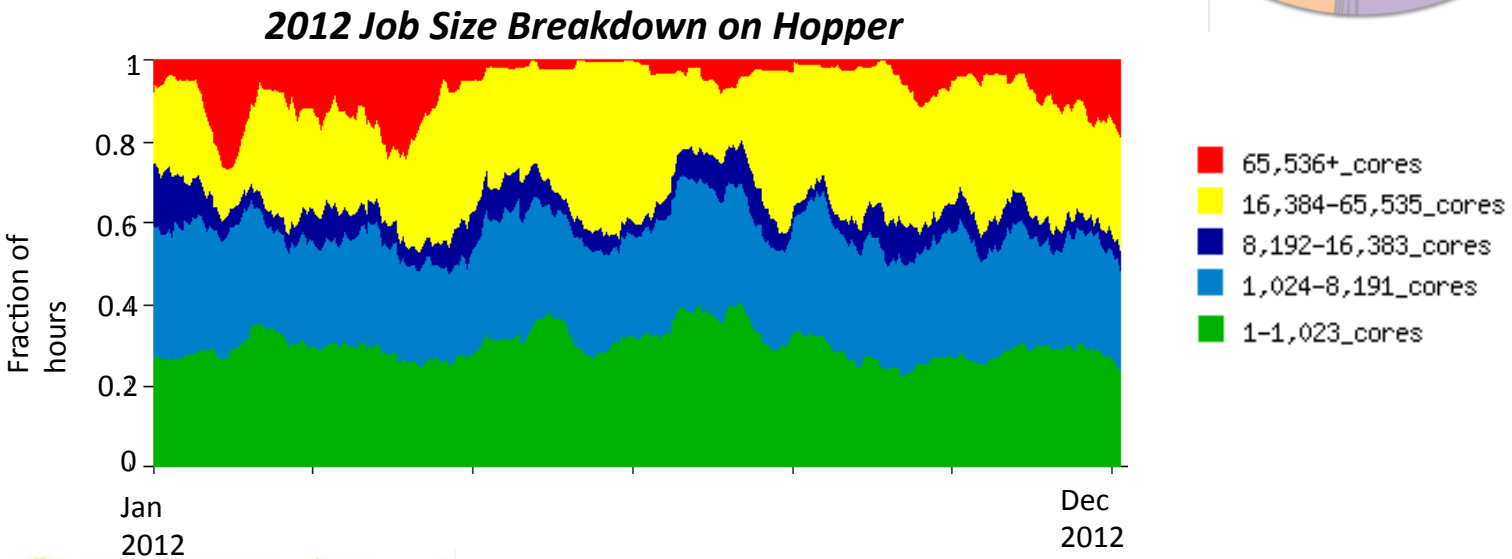
- **4500 users, and we typically add 350 per year**
- **Geographically distributed: 47 states as well as multinational projects**



We support a diverse workload

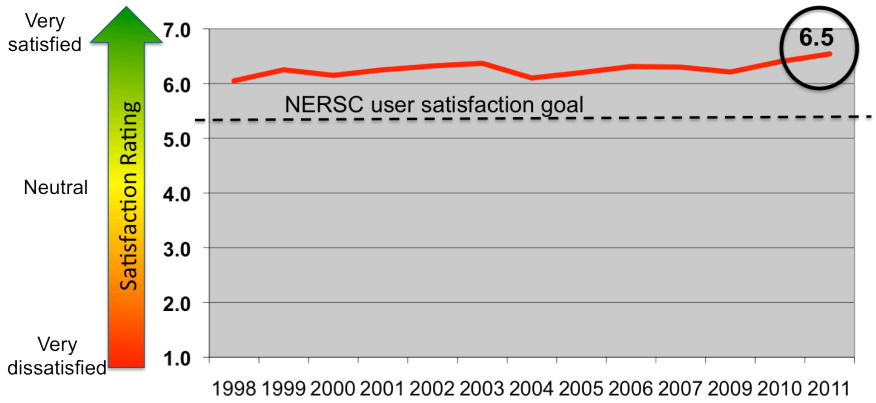
- **Many codes (600+) and algorithms**
- **Computing at scale and at high volume**

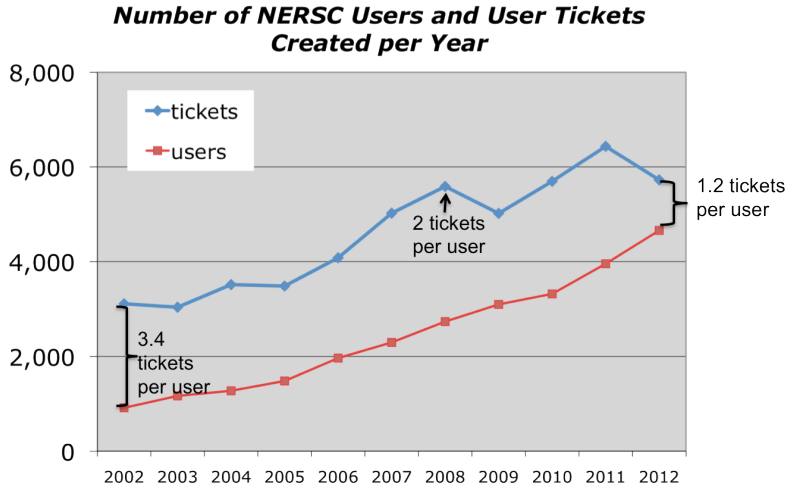




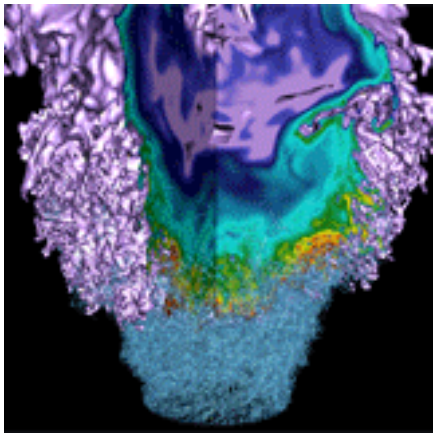
Our operational priority is providing highly available HPC resources backed by exceptional user support

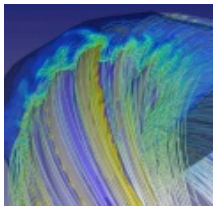
- **We maintain a very high availability of resources (>90%)**
  - One large HPC system is available at all times to run large-scale simulations and solve high throughput problems
- **Our goal is to maximize the productivity of our users**
  - One-on-one consulting
  - Training (e.g., webinars)
  - Extensive use of web pages
  - We solve or have a path to solve 80% of user tickets within three business days

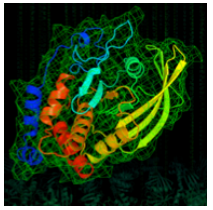


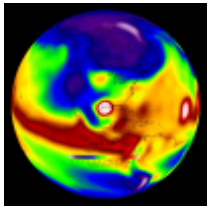


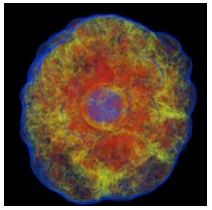
NERSC Today

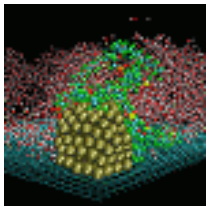




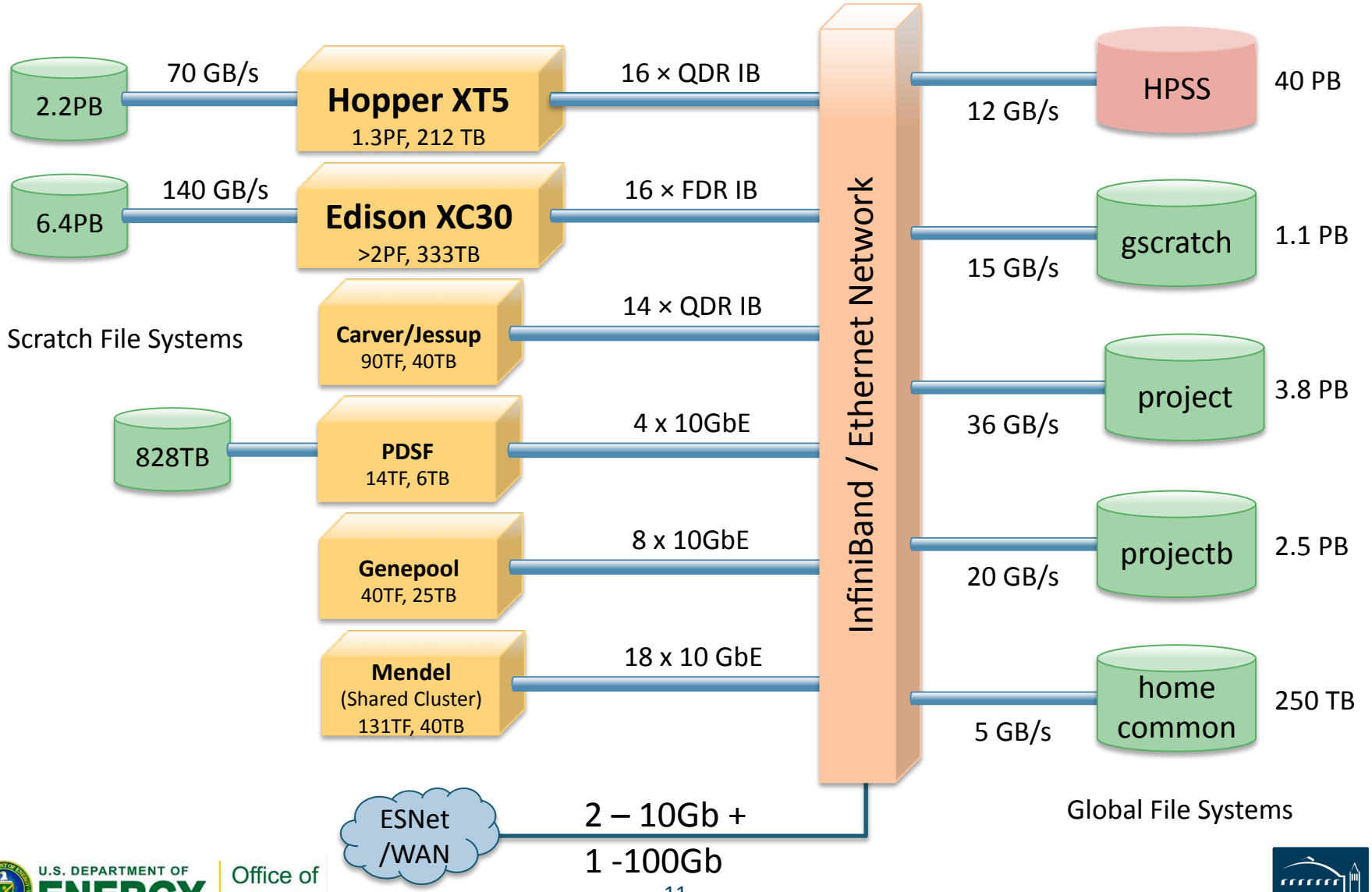








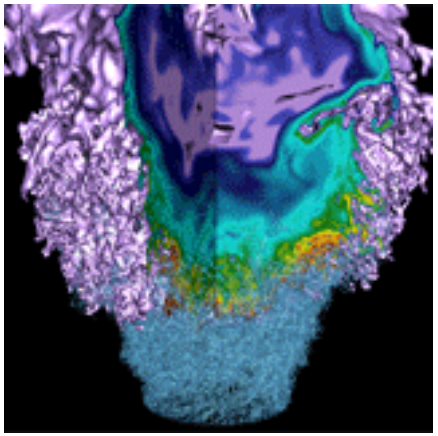
NERSC Systems

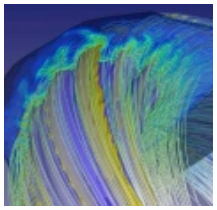


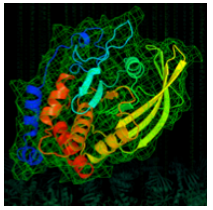
	Hopper	Edison	Mira	Titan
Peak Flops (PF)	1.29	>2.2	10.0	5.26 (CPU) 21.8 (GPU)
CPU cores	152,408	>100,000	786,432	299,008 (CPU) 18,688 (GPU's)
Frequency (GHz)	2.1	2.4	1.6	2.2 (CPU) 0.7 (GPU)
Memory (TB)	217	333	786	598 (CPU) 112 (GPU)
Memory/node (GB)	32	64	16	32 (CPU) 6 (GPU)
Memory BW* (TB/s)	331	442	1406	614 (CPU) 3,270 (GPU)
Memory BW/node* (GB/s)	52	85	29	33 (CPU) 175 (GPU)
Filesystem	2 PB 70 GB/s	6.4 PB 140 GB/s	35 PB 240 GB/s	10 PB 240 GB/s
Sq ft	1956	1200	~1500	4352
Power (MW Linpack)	2.91	2.10	3.95	8.21

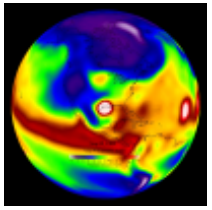
* STREAM

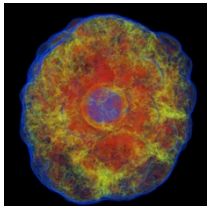
Forecasting

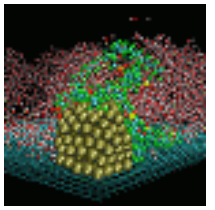






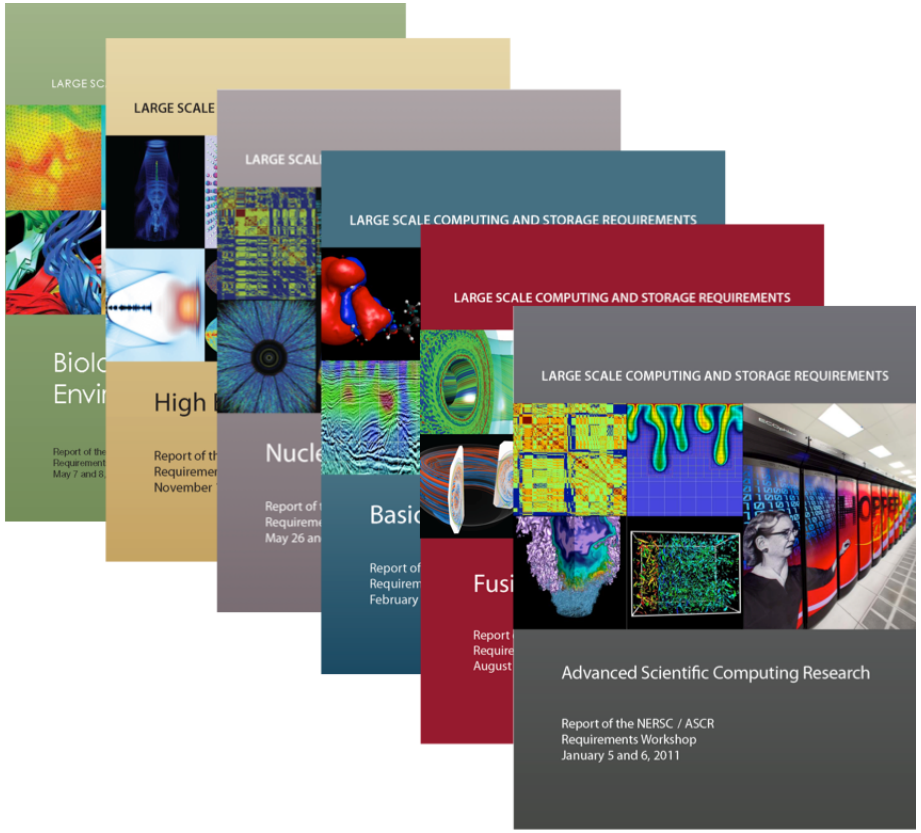






Requirements with six program offices

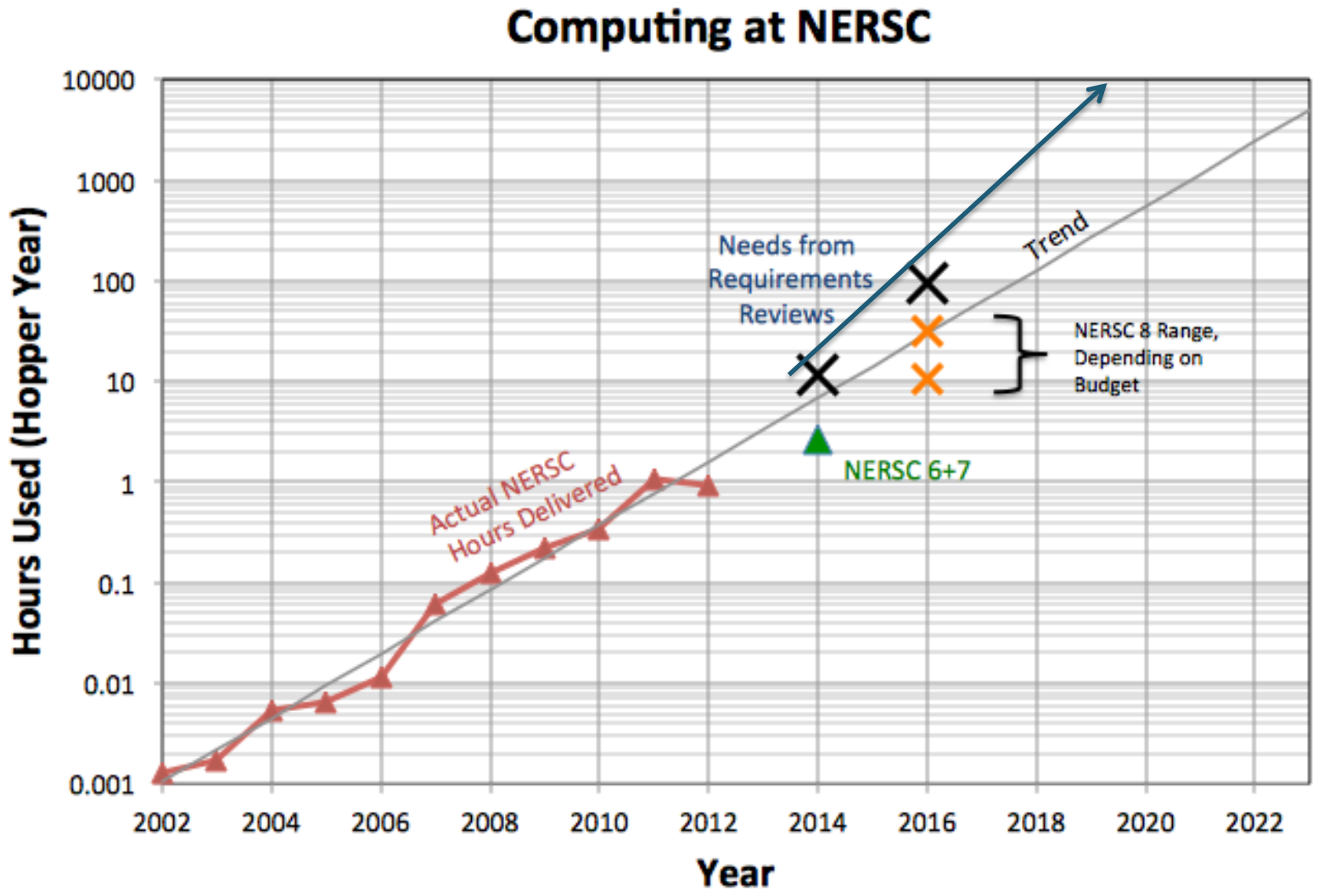
- Reviews with six program offices every three years
- Program managers invite representative set of users (typically represent >50% of usage)
- Identify science goals and representative use cases
- Based on use cases, work with users to estimate requirements
- Re-scale estimates to account for users not at the meeting (based on current usage)
- Aggregate results across the six offices
- Validate against information from in-depth collaborations, NERSC User Group meetings, user surveys



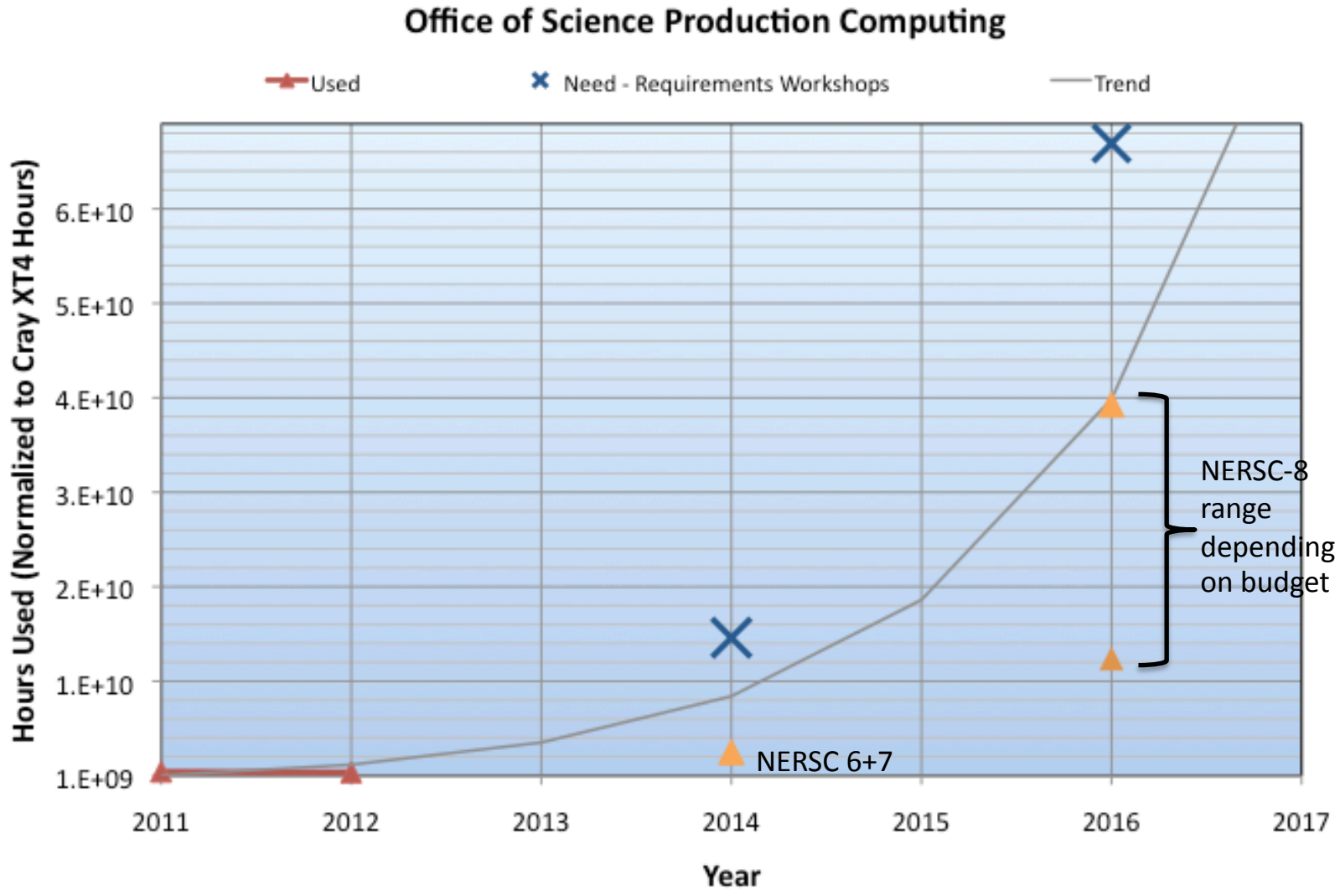
Tends to underestimate need because we are missing future users

http://www.nersc.gov/science/requirements-reviews/final-reports/

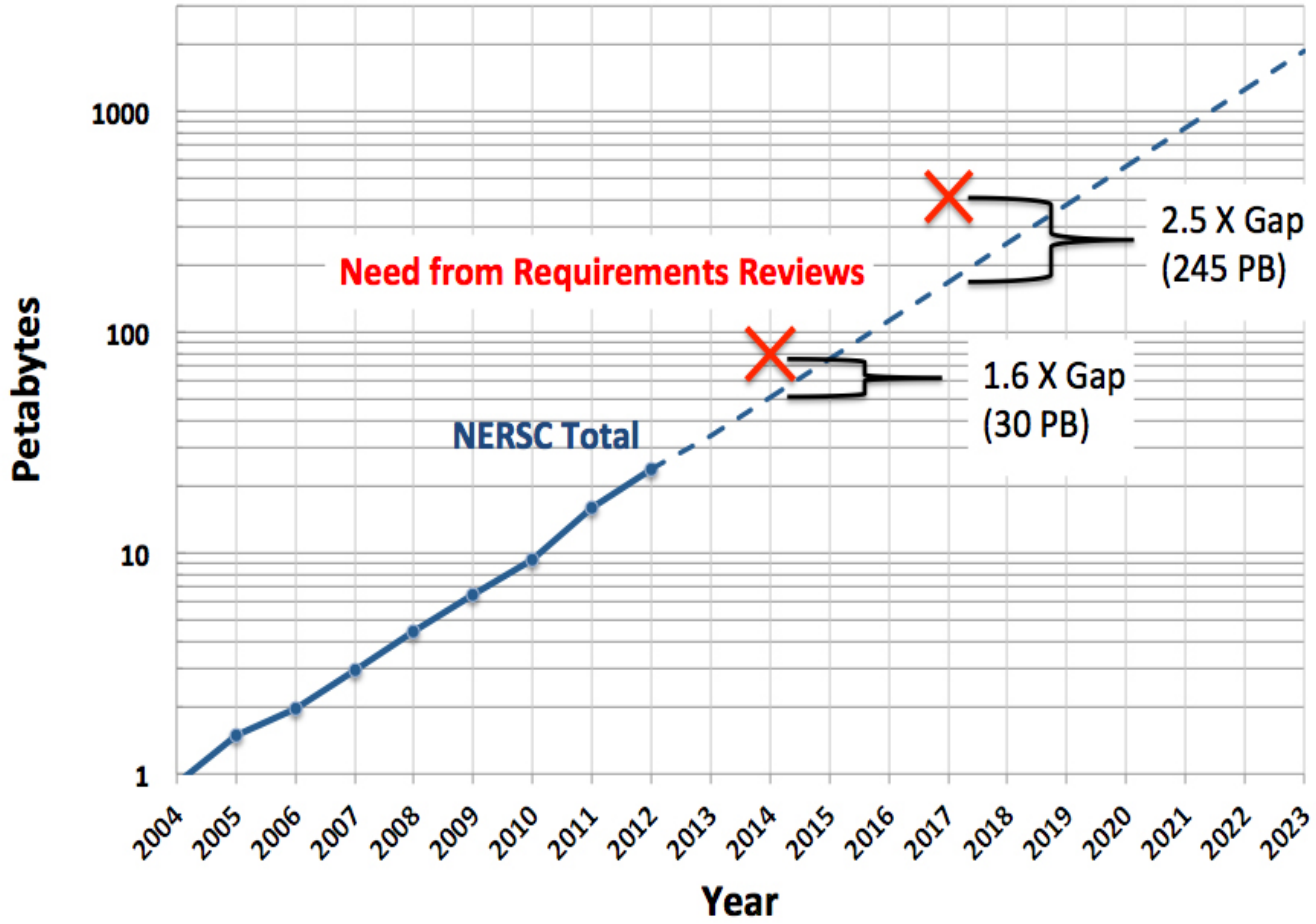
Keeping up with user needs will be a challenge



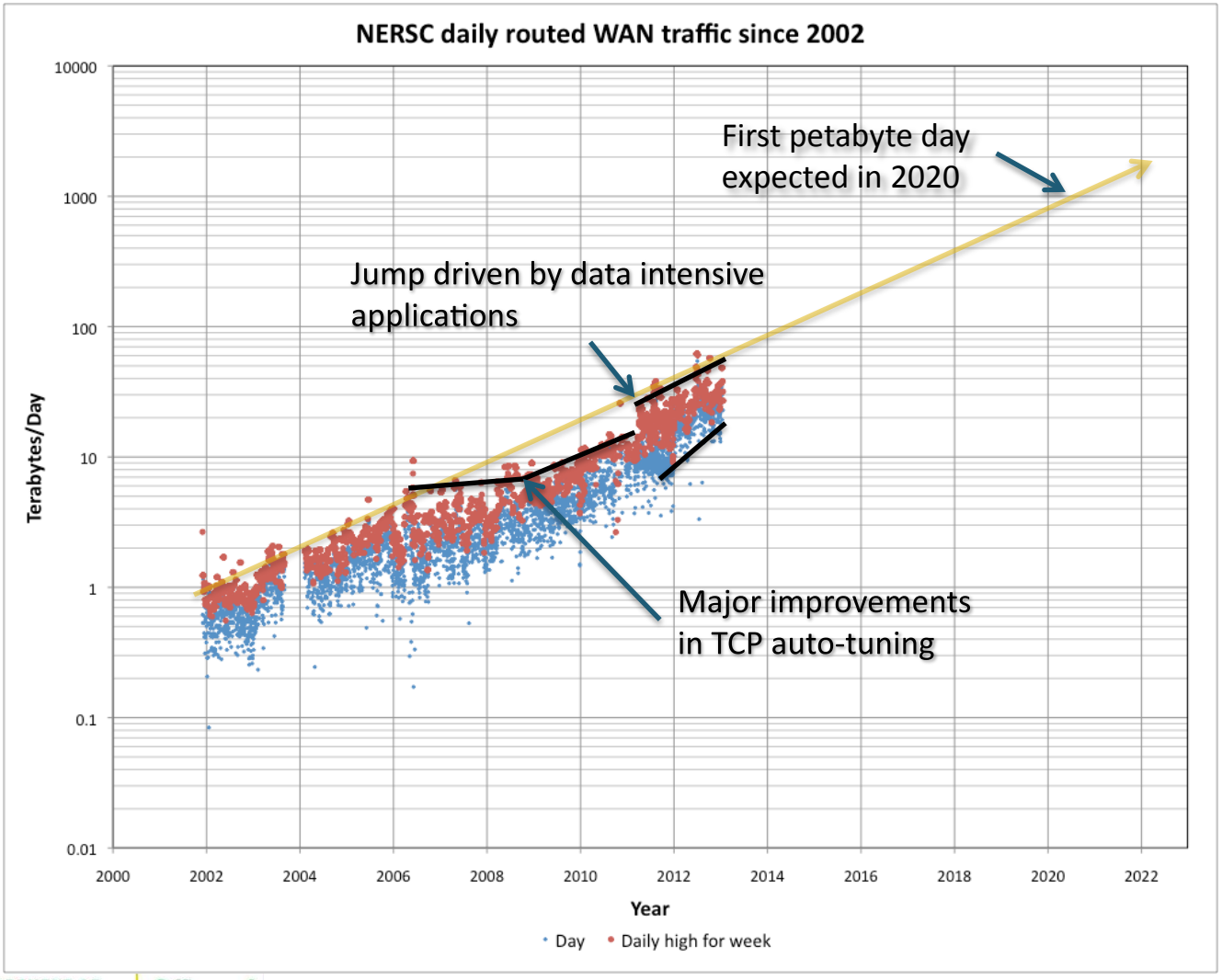
Keeping up with user needs will be a challenge (cont.)



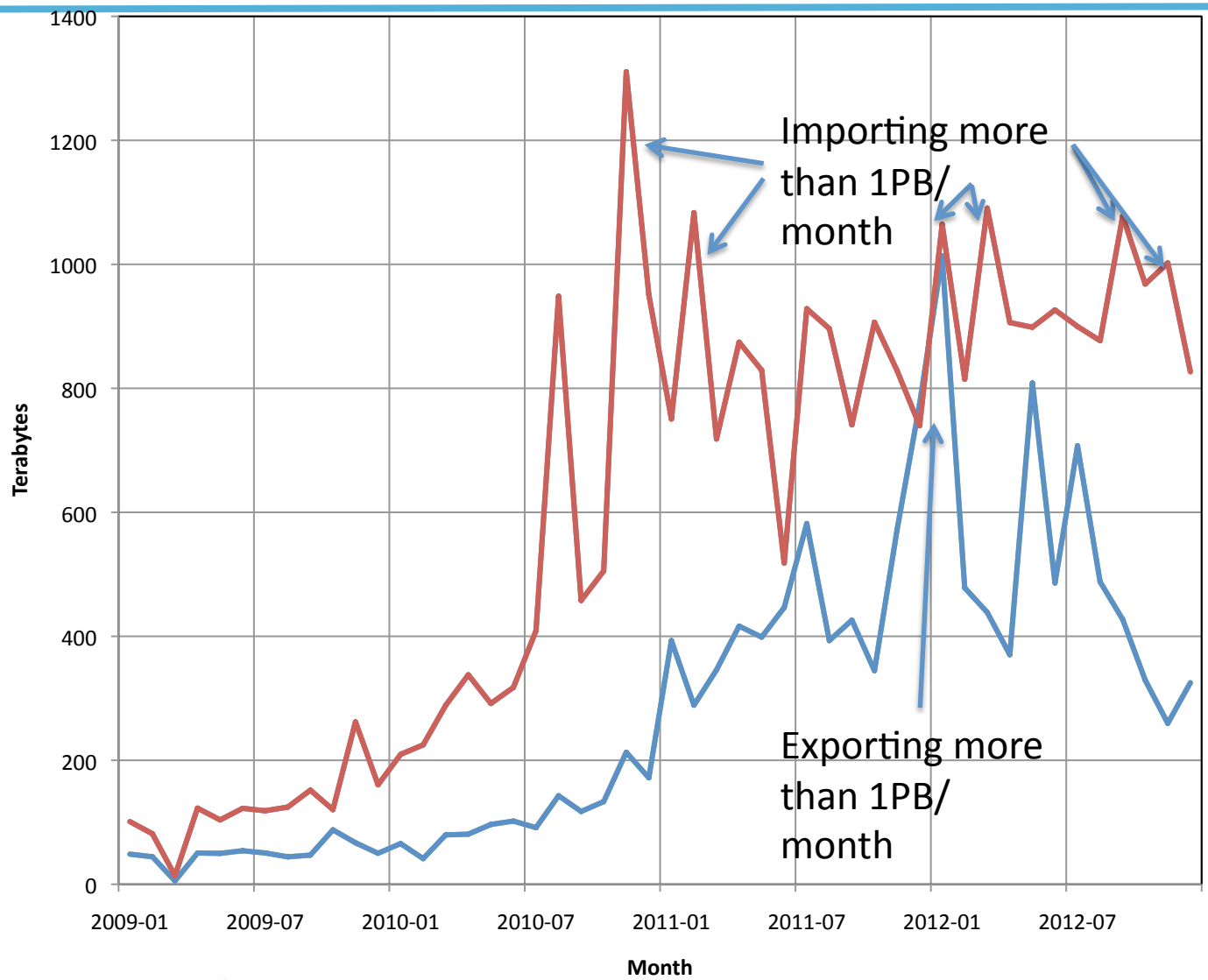
Future archival storage needs



Exponentially increasing data traffic



NERSC users import more data than they export!

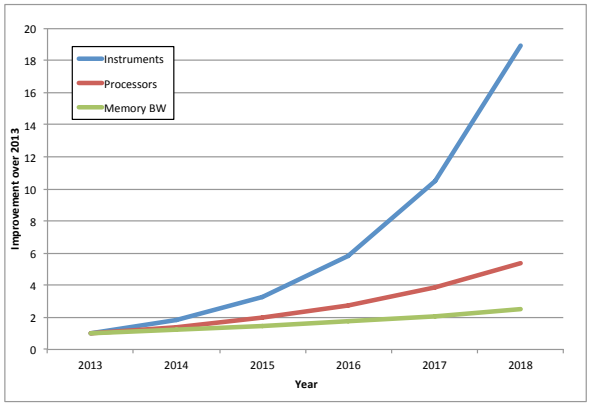


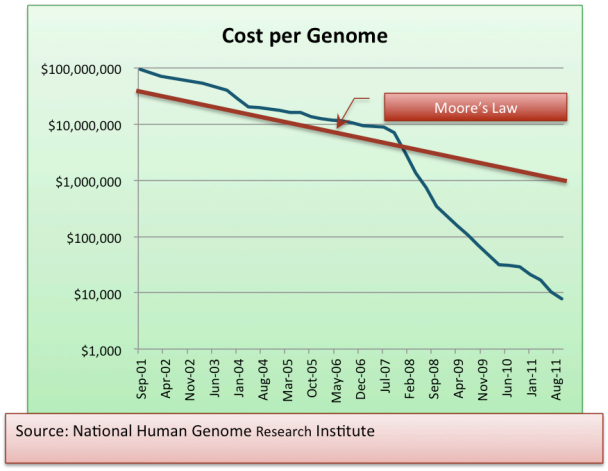
Increased data emphasis in requirements reviews

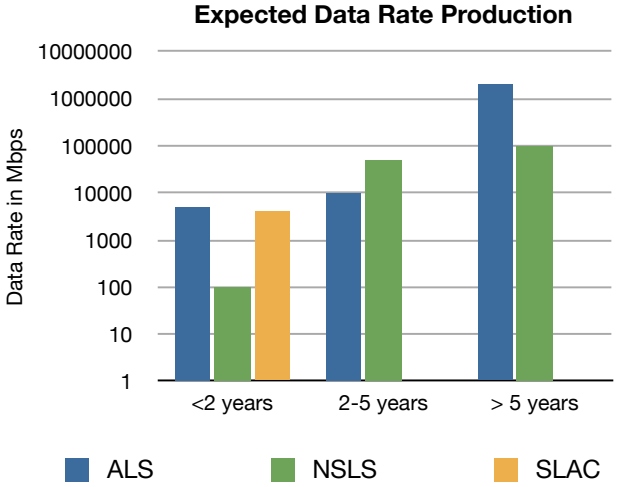
- **BER (2017 draft):** "*Access to more computational and storage resources … and the ability to access, read, and write data at a rate far beyond that available today*"
- **HEP (2017 pre-draft):** "*Need for more computing cycles and fast-access storage; support for data-intensive science, including*
  - *Improvements to archival storage*
  - *Analytics (parallel, DBs, services, gateways etc.)*
  - *Sharing, curation, provenance of data*
- **ASCR (2014): "***Applications will need to be able to read, write, and store 100s of terabytes of data for each simulation run. Many petabytes of long-term storage will be required to store and share data with the scientific community.***"**
- **BES (2014): "***[There is a need to support] … huge volumes of data from the ramp-up of the SLAC LINAC Coherent Light Source (LCLS) [and other experimental facilities in BES]."*
- **FES (2014): "***[Researchers need] data storage systems that can support high-volume/high-throughput I/O.***"**
- **NP (2014):** *Needs include*
  - *"Useable methods for cross-correlating across large databases …"*
  - *"[…] grid infrastructure, including the Open Science Grid (OSG) interface […]. "*
  - *[…] The increased capacity afforded by GPUs has resulted in […] a significant increase in IO demands in both intermediate and long term storage. "*

DOE experimental facilities are also facing extreme data challenges

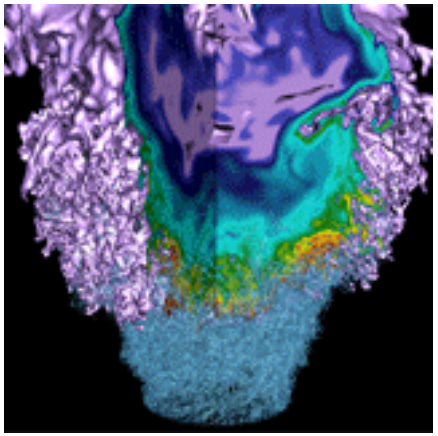
- **The observational dataset for the Large Synoptic Survey Telescope will be ~100 PB**
- **The Daya Bay project will require simulations which will use over 128 PB of aggregate memory**
- **By 2017 ATLAS/CMS will have generated 190 PB**
- **Light Source Data Projections:**
  - 2009: 65 TB/yr
  - 2011: 312 TB/yr
  - 2013: 1.9 PB /yr
  - EB in 2021?
  - NGLS is expected to generate data at a terabit per second

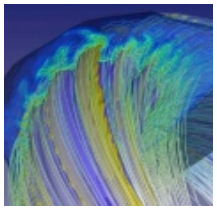


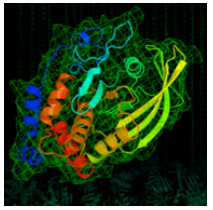


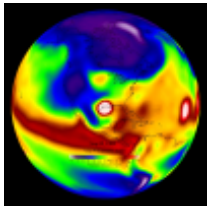


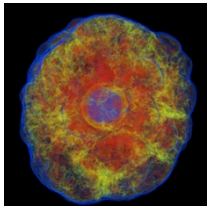
Computing Challenges

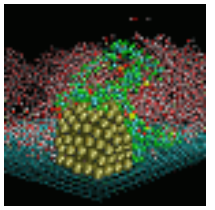












Laws of Physics will Halt Moore's Law

High-performance Logic Technology Requirements (ITRS 2011)

Year	2012	2013	2014	2015	2016	2017	2018	2019	2020
Gate Length	22	20	18	17	15.3	14	12.8	11.7	10.6
Equivalent Oxide Thickness	●	●	●	●	●	●	●	●	●
Source-Drain Leakage	●	●	●	●	●	●	●	●	●
Threshold Voltage	●	●	●	●	●	●	●	●	●
CV/I Intrinsic Delay	●	●	●	●	●	●	●	●	●
Total Gate Capacitance	●	●	●	●	●	●	●	●	●
Drive Current	●	●	●	●	●	●	●	●	●

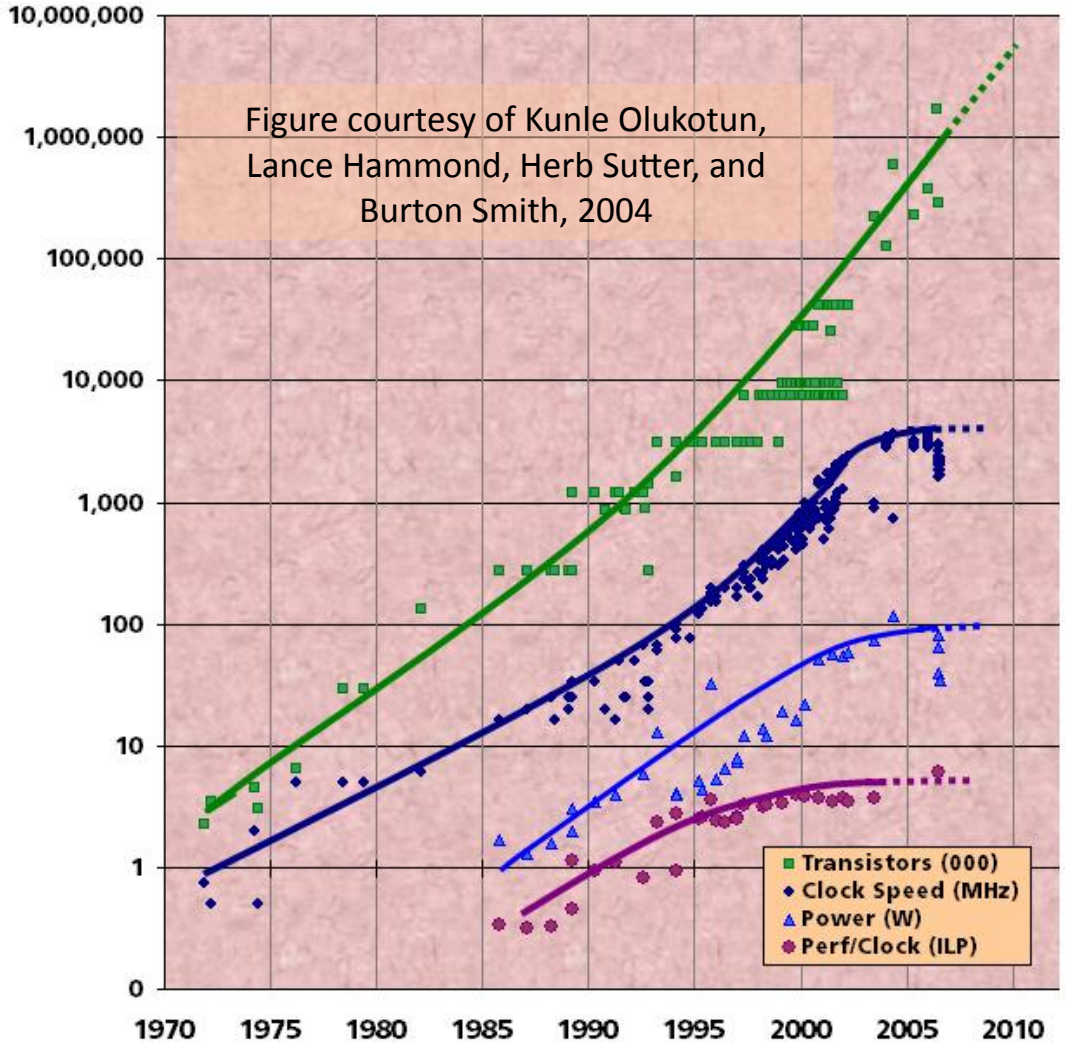
- **Time line shown for best performing multi-gate transistor technology.**
- **Similar timelines exist for other functional components; *e.g.*, memory, RF logic.**

● technology available
● solutions known
● no known solutions

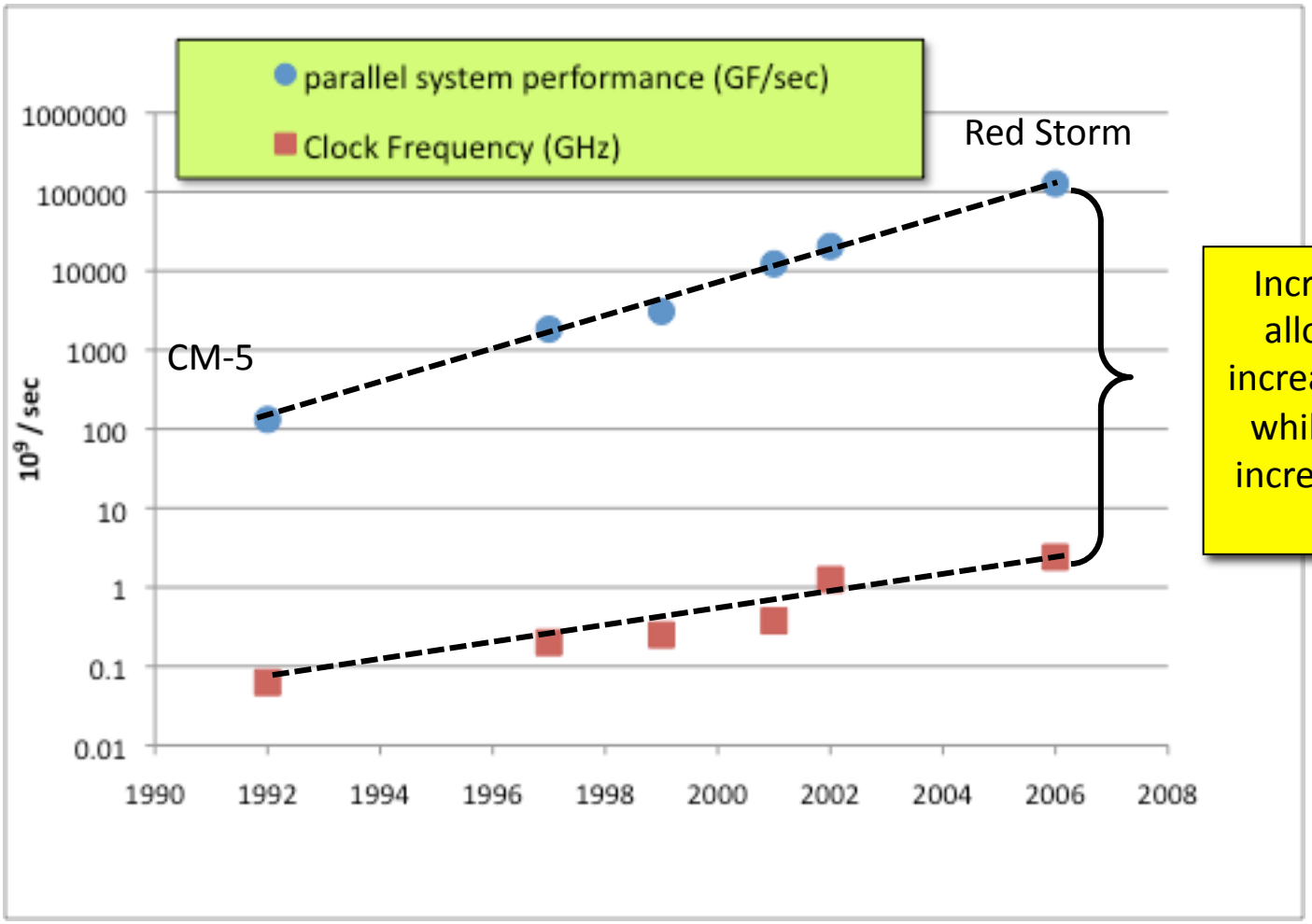
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Clock speeds are expected to stay near 1 GHz

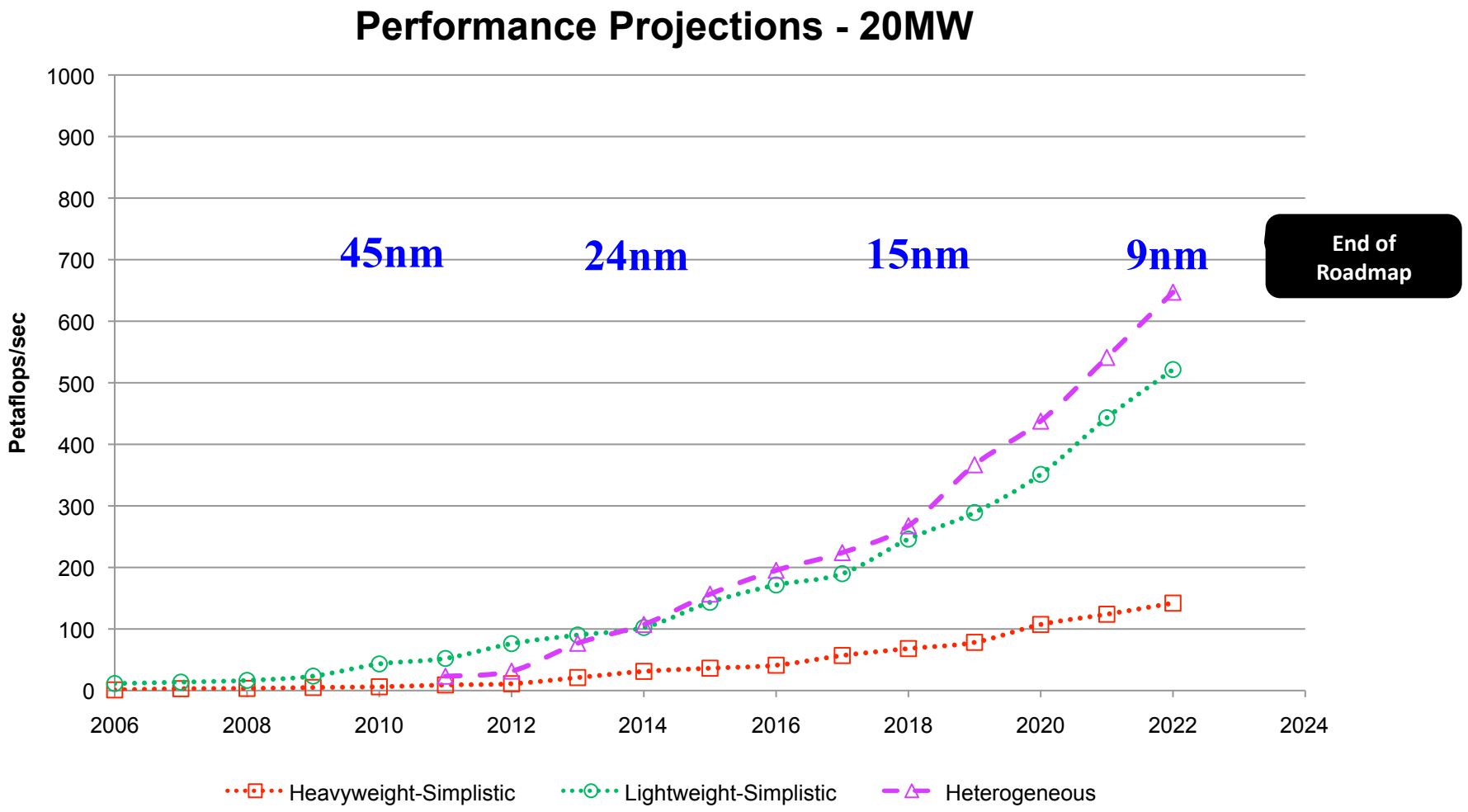


Concurrency is one key ingredient in getting to exaflop/sec

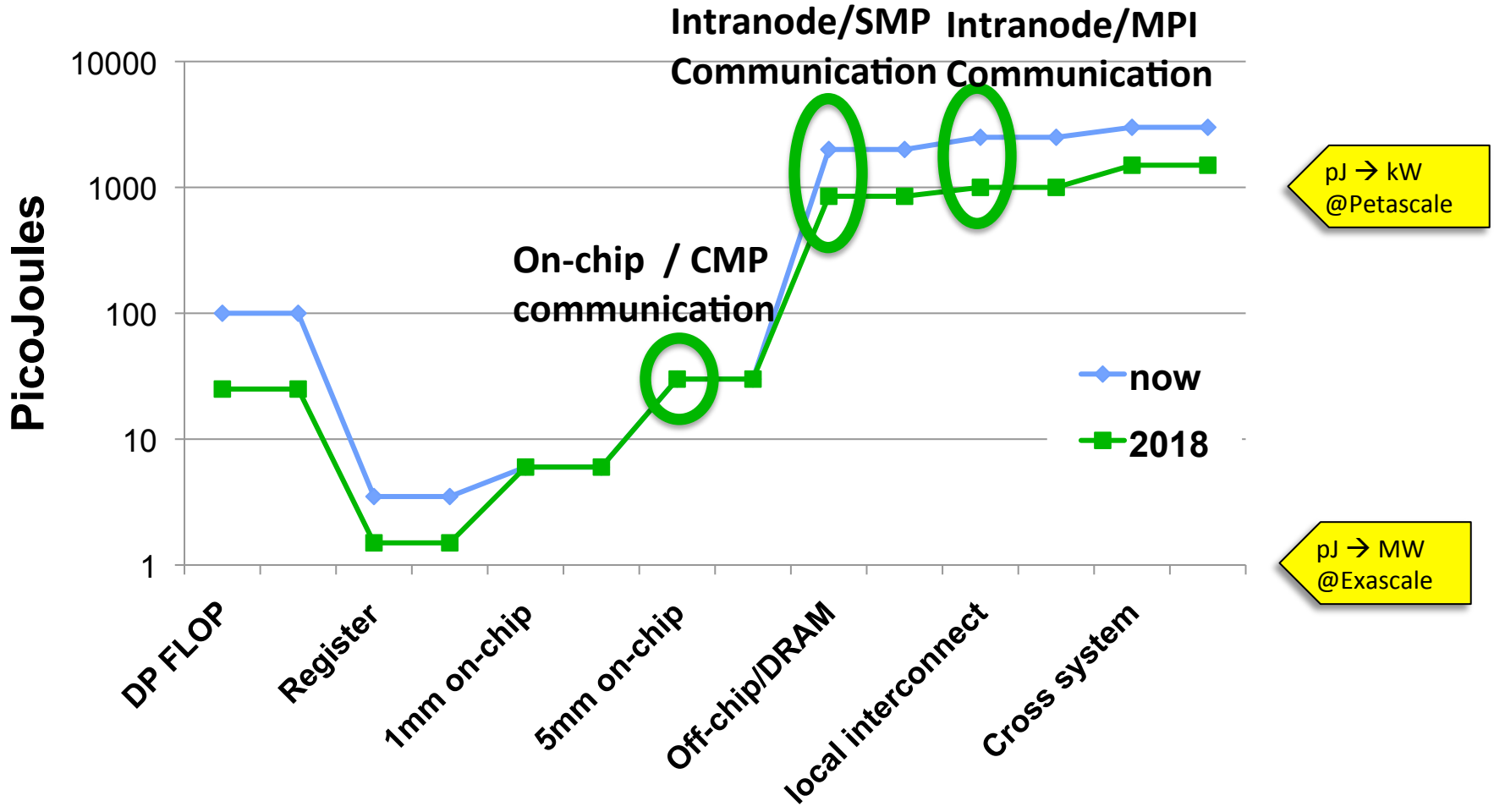


Increased parallelism allowed a 1000-fold increase in performance while the clock speed increased by a factor of 40

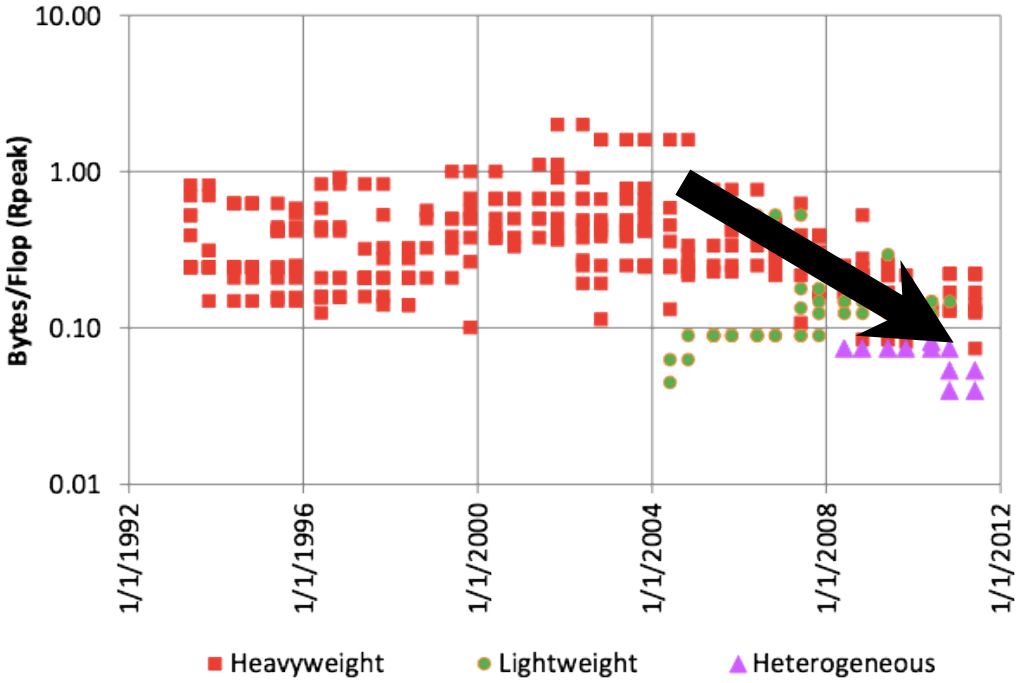
Future gains in supercomputing will be limited by power

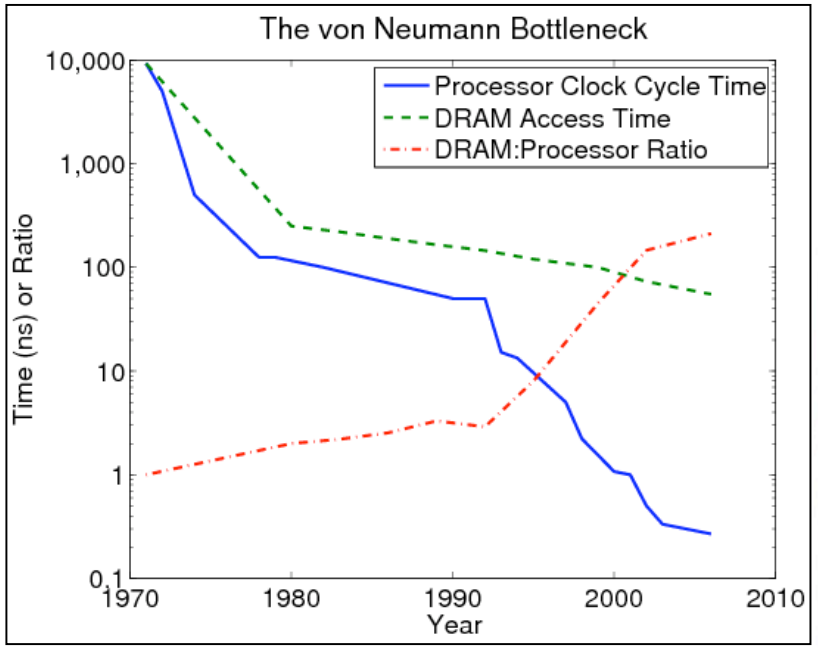


Where does the energy go?



Both memory capacity and bandwidth are significant issues for DOE applications

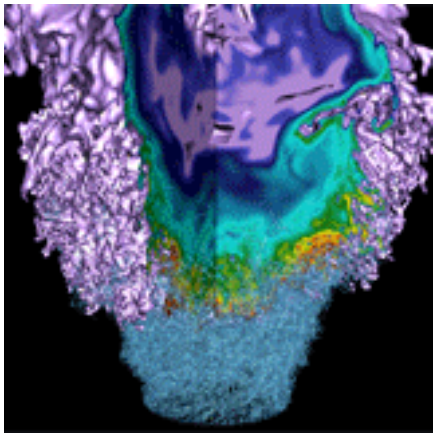


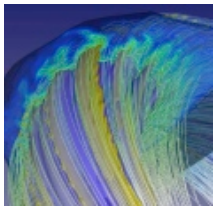


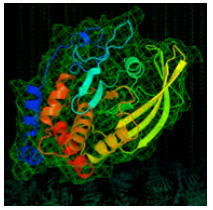
NNSA National Nuclear Security Administration

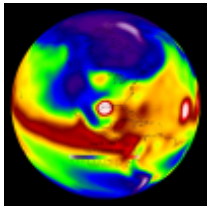
Sandia National Laboratories

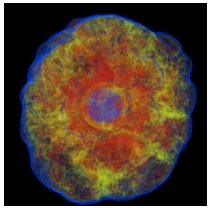
NERSC Strategy

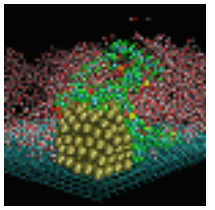










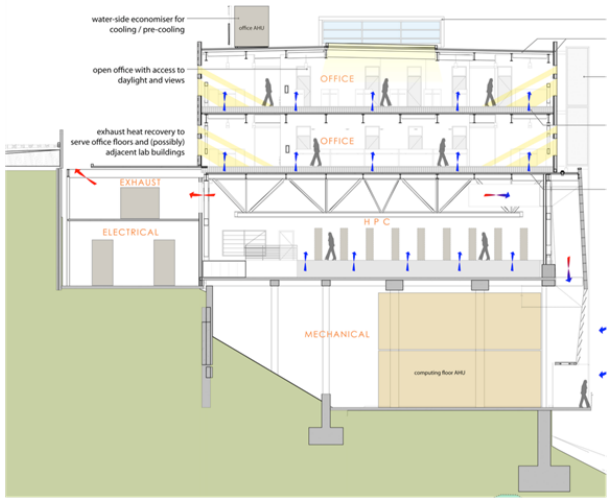


Strategic Objectives

- **Meet the ever-growing computing and data needs of our users by**
  - providing usable exascale computing and storage systems
  - transitioning SC codes to execute effectively on manycore architectures
  - influencing the computer industry to ensure that future systems meet the mission needs of SC
- **Increase the productivity, usability, and impact of DOE's user facilities by providing comprehensive data systems and services to store, analyze, manage, and share data from those facilities**

We are deploying the CRT facility to meet the ever- growing computing and data needs of our users

- **Four story, 140,000 GSF**
  - Two 20 Ksf office floors, 300 offices
  - 20 K -> 29 Ksf HPC floor
  - Mechanical floor
- **42 MW to building**
  - 12.5 MW initially provisioned
  - WAPA power: Green hydro
- **Energy efficient**
  - Year-round free air and water cooling
  - PUE < 1.1
  - LEED Gold
- **Occupancy Early 2015**



Providing usable exascale computing and storage systems

- **We made NERSC-7 an x86-based system because our broad user base wasn't ready in 2013 for GPUs, accelerators or greatly increased threading**
- **We will deploy pre-exascale systems in 2016 (NERSC-8) and 2019 (NERSC-9), and an exascale system in 2022. Our strategy is:**
  - Open competition for best solutions
  - Focus on the performance of a broad range of applications, not synthetic benchmarks
  - General-purpose architectures are needed in order to support a wide range of applications, both large-scale simulations and high volumes of smaller simulations
  - Earlier procurements to influence designs
  - Leverage Fast Forward and Design Forward
  - Engage co-design efforts
  - Transition users to a new programming model

NEW

Programming Models Strategy

- **The necessary characteristics for broad adoption of a new pmodel is**
  - Performance: At least 10x-50x performance improvement
  - Portability: Code performs well on multiple platforms
  - Durability: Solution must be good for a decade or more
  - Availability/Ubiquity: Cannot be a proprietary solution
- **Our near-term strategy is**
  - Smooth progression to exascale from a user's point of view
  - Support for legacy code, albeit at less than optimal performance
  - Reasonable performance with MPI+OpenMP
  - Support for a variety of programming models
  - Support optimized libraries

Strategy for Transitioning the SC Workload to Energy Efficient Architectures

- We will deploy testbeds to gain experience with new technologies and to better understand emerging programming models and potential tradeoffs.
- We will have in-depth collaborations with selected users and application teams to begin transitioning their codes to our testbeds and to NERSC-8
- We will develop training and online resources to help the rest of our users based on our in-depth collaborations, as well as on results from co-design centers and ASCR research
- We will add consultants with an algorithms background who can help users when they have questions about improving the performance of key code kernels

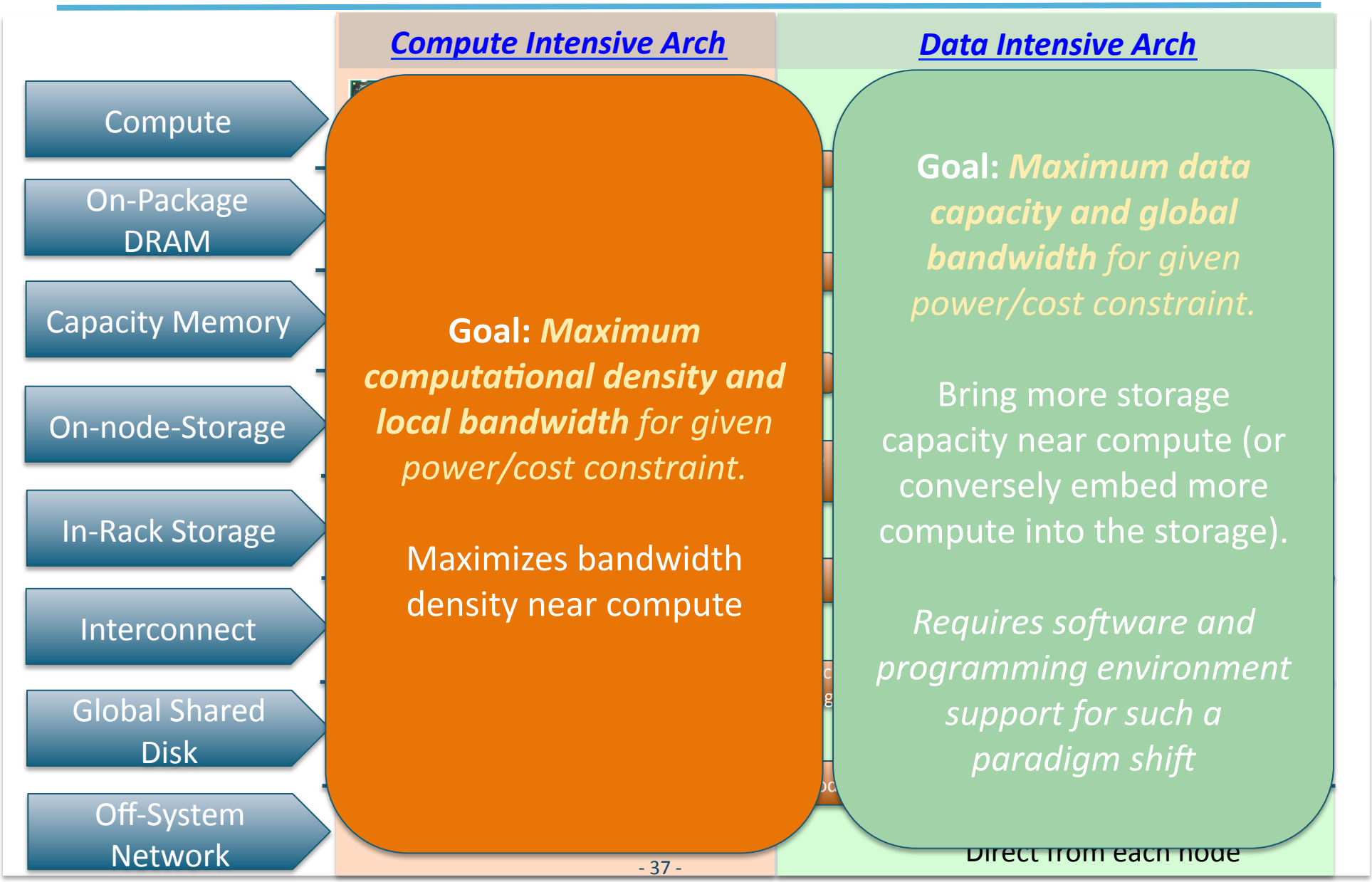
Strategy for ensuring that future systems meet SC mission requirements

- Partner with Los Alamos and Sandia on procurements in 2016 and 2019. The larger size of these procurements will give us greater leverage with industry
- Provide industry with greater information on NERSC's workload through new and innovative instrumentation, measurement, and analysis
- Actively engage with industry through DOE's Fast Forward and Design Forward programs
- Leverage the Berkeley/Sandia Computer Architecture Laboratory (CAL) that has been established by ASCR
- Serve as a conduit for information flow between computer companies and our user community

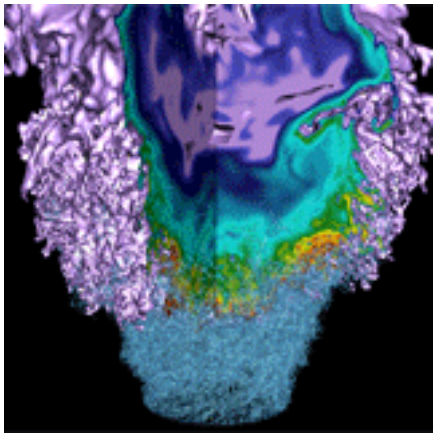
Extreme Data Strategy

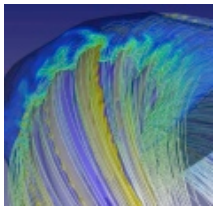
- Partner with DOE experimental facilities to identify requirements and create early success
- Develop and deploy new data resources and capabilities
- Provide new classes of HPC expertise required for data-intensive workloads
- Leverage ESnet and ASCR research to create end-to-end solutions

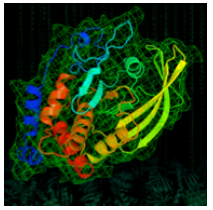
Unique data-centric resources will be needed

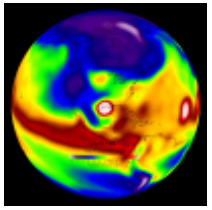


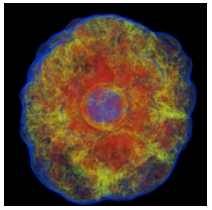
NERSC System Plan

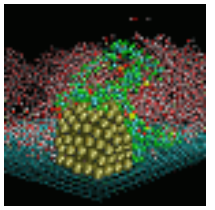












Projections of Installed Capacity